FINAL REPORT

Title: "Design of an Intelligent Machine"

Grant number: ID#: NAG9-736

Principal Investigator: Dr. G. G. Johnson

Start Date : September 1, 1994
Duration : 12 months
Extension : 6 months

University of Houston
College of Natural Science and Mathematics
Department of Mathematics
4800 Calhoun Blvd.
Houston, Texas 77204

Final Report:

The work originally intended was diverted by the new problem of finding a method to translate documents written in Russian, to English. There were in 1995, no commercially available programs to perform this task. Several programs in the development stage were found. The prototypes were tried on crisply printed copy. They were moderately successful in dealing with straight text. They were quick, in that a page was processed and within three to ten seconds the translated text was displayed. The accuracy was about ninety five percent. But all had the problem of dealing with phrases that were within rectangular boxes. Such displays are quite common in technical documents, and often there are several consecutive pages of such displays with little additional text to assist in understanding the content of such displays. There were no trials using fax copy, which is usually found to have characters somewhat blurred.

There is a need to find a quick, though not necessarily extremely accurate, way to translate such documents which have such displays. The reason is, that having determined roughly the content, if it is deemed important, then an accurate translation could be done by a professional translator.

What was discovered is that the machine translation of the determined characters was not the problem, but rather the determining of the digitized characters of the text. The digitizing is usually carried out at 200 dots per inch (dpi). This is quite satisfactory for straight text that is crisply printed. The problem comes when fax copy is processed or when there are words encased within rectangular boxes. In the fax case, the characters are often blurred together, while in box case the characters often intersect the border, or are very close to the border. In these cases the search routines used to determine the characters from the digitized data are unable to determine the letters displayed. The problem becomes quite severe when words are underlined, or encased within the body of the text. Increasing the digitizing to 300 dpi does not alleviate the

problem of characters touching lines nor the problem of blurred characters.

Having found the cause of some of the difficulties that could not resolved by the prototypes presently available, the work then turned to finding methods to resolve them.

Printed text that included diagrams and encased phrases were digitized at 150 dpi to 300 dpi and stored. Programs were then constructed to process the stored data. The first were used to isolate the encased or underline words and characters from the remainder of the text, allowing the portions in which the words with characters that could be recognized correctly to be translated.

There was then the problem of filtering the blurred, encased or underlined characters to separate them from one another as well as from the boxes or lines, without removing the distinct information that determine characters.

The most severe difficulties were found in the blurred word case. several techniques such as WRRM, were tried. These did not yield usable results. The problem remains unsolved.

There is much that can yet be tried, particularly in the underlined and encased words.